

Лабораторная работа по анализу текстов

Можно ли различить голоса персонажей?

Стилометрическая классификация 34 персонажей Stardew Valley по 2005 репликам

Три блока · три рассказчика

Корпус → методология → результаты

2005 реплик · 34 персонажа · один автор · 5-блочная перекрёстная проверка

Структура доклада

01

Корпус

Рассказчик 1

- 2005 реплик
- 34 персонажа
- сбор, очистка, разметка
- ограничения данных

02

Методология

Рассказчик 2

- признаки и LSA
- классификаторы
- перекрёстная проверка
- метрики

03

Результаты

Рассказчик 3

- лучшие модели
- матрицы ошибок
- кого путает модель
- выводы

Блок 1

Корпус: как собрали и разметили

Что именно мы классифицируем и почему это сложная задача

- Откуда данные и какова единица анализа
- Кто в корпусе и как распределены реплики
- Длина реплик и очистка
- Паспорт корпуса и ограничения

Рассказчик 1

Что анализируем

Постановка задачи

- Единица анализа — одна реплика персонажа.
- Всего 2005 реплик; 34 персонажа; один автор.
- Класс — имя персонажа; один объект относится ровно к одному классу.
- Реплики короткие: в среднем 19 слов, медиана 17.
- Корпус англоязычный; разметка — только текст и персонаж.

Примеры разных голосов

Abigail (139 реплик)

«Oh, that's right... I heard someone new was moving onto that old farm.»

Jas (34 реплики)

«Wow, a birthday present! I love it!»

Krobus (28 реплик)

«I wanted to give you this as a 'thank you' for letting me live here with you.»

Wizard (24 реплики)

«Ah, yes. I have predicted your arrival a long time ago, young PLAYER.»

Кто в корпусе

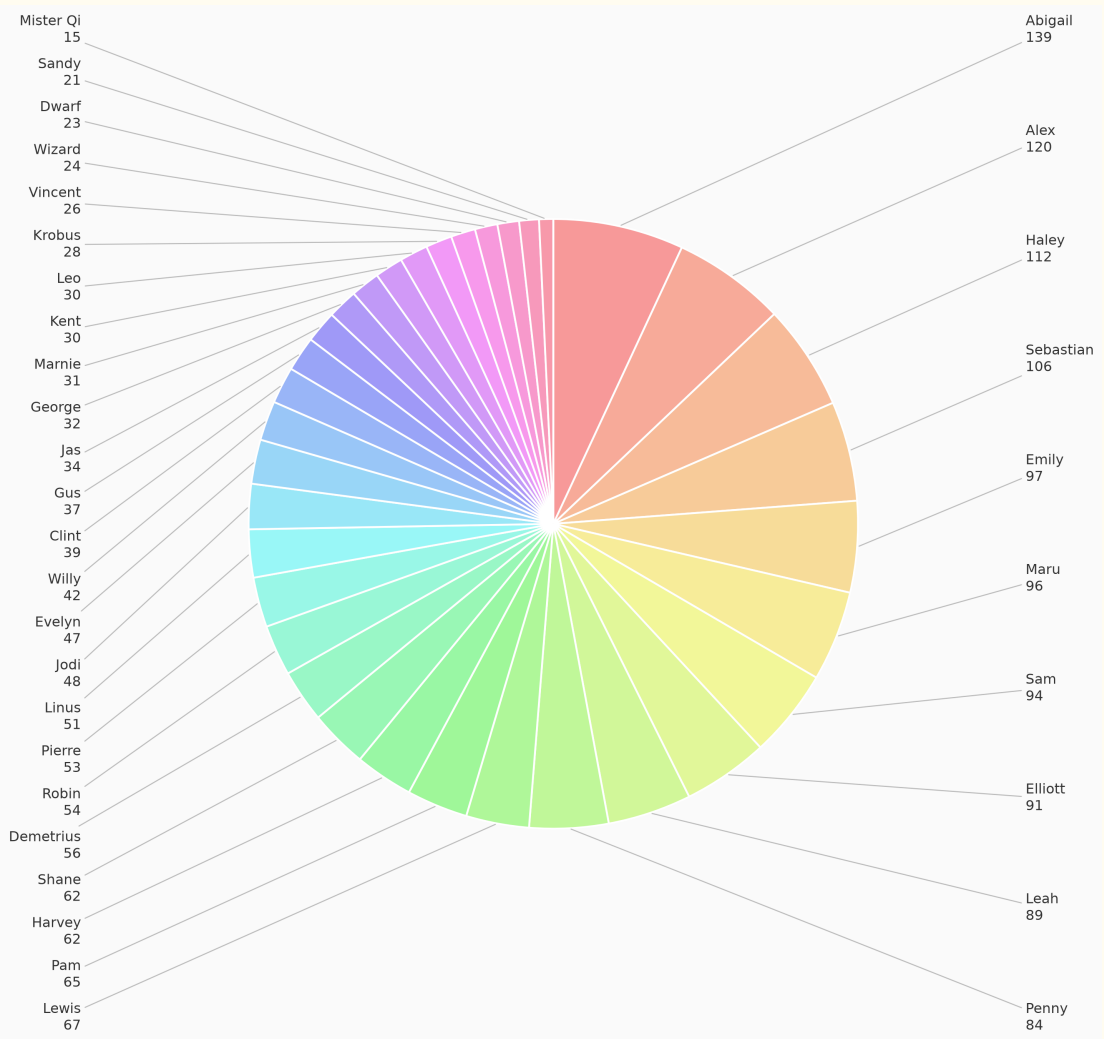


Состав

- 34 персонажа: подростки, взрослые, пожилые, дети и нелюди.
- Все реплики написаны одним разработчиком; многие — в одних и тех же обновлениях и сюжетных контекстах.
- Классы сильно неравномерны: от 15 до 139 реплик.
- Метаданные о поле, возрасте и виде используются только для анализа ошибок, не как признаки.

Сколько реплик у каждого

Радиальная диаграмма: число реплик по персонажам

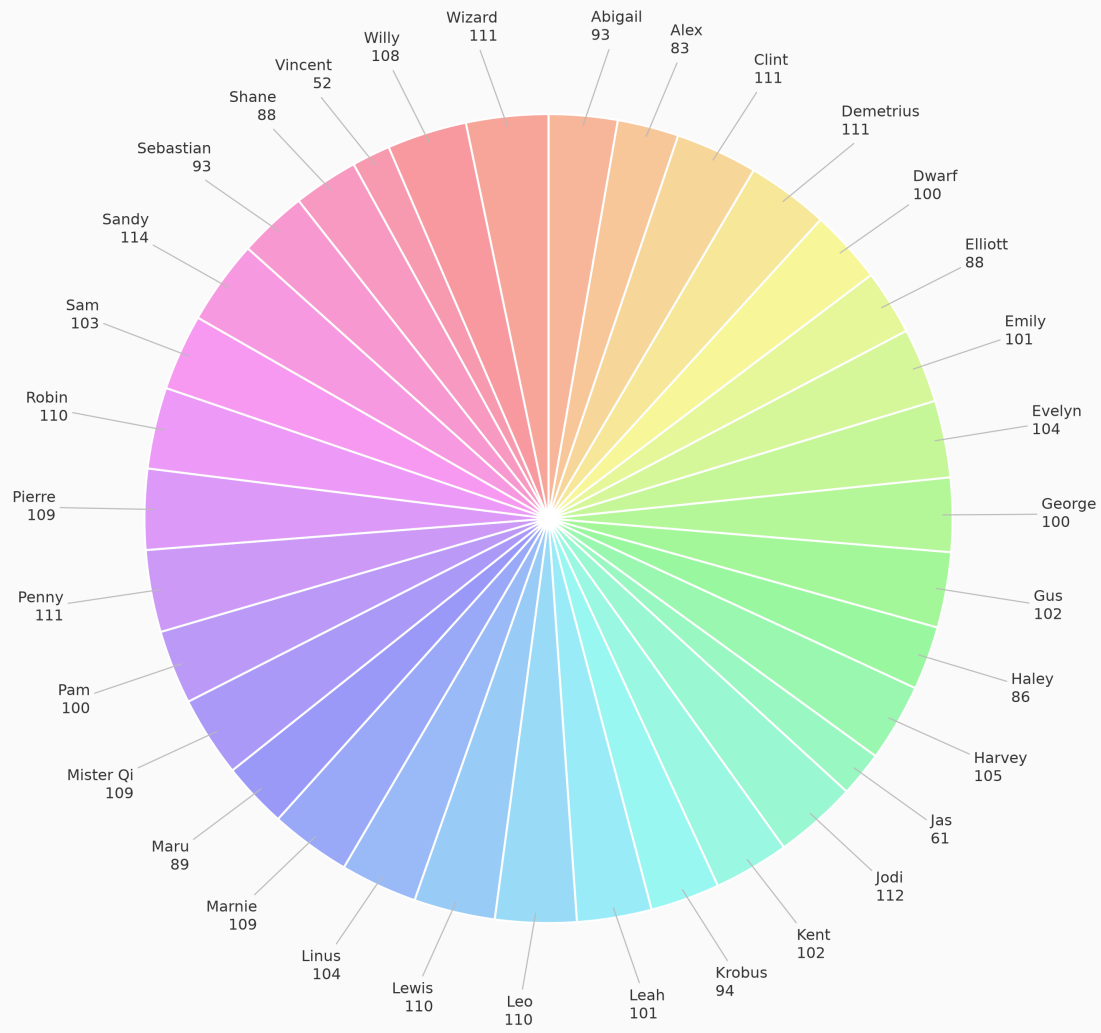


Дисбаланс классов

- Максимум — Abigail: 139 реплик; минимум — Mister Qi: 15.
- Разница между классами почти десятикратная.
- При 34 классах случайное угадывание даёт около 2,9%.
- Поэтому доля правильных ответов обманчива: модель может выбирать частые классы.
- Основная метрика — макро-усреднённая F-мера: каждый персонаж весит одинаково.

Длина реплик

Радиальная диаграмма: медианная длина предложения в символах



Короткие тексты

- Средняя длина — 19,1 слова; медиана — 17 слов.
- Диапазон — от 3 до 78 слов.
- Медианная длина предложения в символах: от 52 у Vincent до 114 у Sandy.
- Короткие реплики дают мало свидетельств; стилометрические оценки на них нестабильны.
- Длина сама по себе — сильный, но нежелательный признак: она может стать прокси персонажа.

Очистка и разметка

Что сделали

• Источник — игровые реплики; CSV в cp1251, XLSX как резерв.

• Нормализация Unicode: многоточия, кавычки, тире, неразрывные пробелы.

• Плейсхолдеры <PLAYER>, <FARM> и другие заменены нейтральными словами; их количество сохранено отдельным признаком.

• Токенизация и разбиение на предложения — NLTK; частеречная разметка — Penn Treebank → Universal POS.

• Стоп-слова, пунктуация и регистр не удалялись: это стилометрический сигнал.

Что считаем шумом

• Шум: дубликаты, технические артефакты, общие короткие фразы.

• Ручная фильтрация уже удалила короткие и общие реплики; правила не сохранились.

• В текущем CSV 0 точных дублей; 13 пар с косинусной близостью ≥ 0,8 — все внутри одного персонажа.

• Нет идентификаторов сцен и диалогов; это ограничение анализа.

• Разметка — только имя персонажа; никакой ручной стилистической разметки нет.

Паспорт корпуса

Показатель	Значение
Реплик	2005
Персонажей	34
Словоупотреблений	38 207
Средняя длина	19,1 слова
Медиана длины	17 слов
Минимум / максимум	3 / 78 слов
Классы	15–139 реплик
Точные дубли	0
Почти-дубли (cos ≥ 0,8)	13 пар, все внутри класса

Что важно

- Корпус пригоден для классификации, но не для сильных выводов о «чистом» стиле.
- Ограничения: короткие тексты, дисбаланс, один автор, отсутствие сцен.
- Все дальнейшие результаты нужно сравнивать с случайным уровнем 2,9%, а не с нулём.

Блок 2

Методология: как измеряем стиль

Признаки, модели и проверка — на понятном языке

- Какие свойства текста считаем
- Как работает LSA
- Как устроены классификаторы
- Как не обмануть себя при оценке

Рассказчик 2

Общая схема

Реплика
текст персонажа

Очистка
Unicode, плейсхолдеры, пробелы

Токены и POS
NLTK, Universal POS

Признаки
ручные, n-граммы, LSA

Модель
вероятности по 34 классам

Проверка
5 блоков, повторы, метрики

- Обучение и преобразования выполняются только внутри обучающих блоков; тест не участвует в настройке.
- Вектор признаков — это не «понимание» текста, а числовое описание его статистических свойств.

Признаки: что именно считаем

Ручные признаки

- 454 ручных признака в 11 группах.
- Длина: слова, символы, предложения, длина слова.
- Лексика и орфография: распределение длин, апострофы, заглавные, цифры.
- Словарное богатство: TTR, Guiraud, Yule's K, hapax; на коротких текстах нестабильно.
- Пунктуация: запятые, многоточия, восклицания, вопросы, повторы.
- Разговорные маркеры: междометия, смех, хеджи, вежливость, сокращения.

Синтаксис и векторные блоки

- Функциональные слова: около 150 закрытых слов и 12 категорий.
- Местоимения: 1-е, 2-е, 3-е лицо; отношение «я/ты».
- Части речи: Universal POS и Penn Treebank, время, модальность, герундий.
- Синтаксис-прокси: клаузы, подчинительные союзы, пассив, вопросы.
- Символьные n-граммы 2–5 и POS n-граммы 1–3.
- LSA: сжатые векторы по словам, содержанию, POS и символам.

LSA простыми словами

Идея

- Строим таблицу «документ × термин» и взвешиваем частоты через TF-IDF.
- Сжимаем таблицу сингулярным разложением до k скрытых осей.
- Каждая реплика превращается в вектор из k чисел.
- Оси не обязаны быть темами; их знак и порядок произвольны.
- k выбирается по объяснённой дисперсии и внешней метрике.

Варианты

- word — лексика и тема.
- content — содержание после удаления стоп-слов.
- fw — функциональные слова и синтаксис.
- char — морфология и пунктуация.
- pos — частеречные последовательности.
- В работе k = 50–400; для content-LSA качество росло до 400.

Ключевой вопрос: LSA по содержательным словам ловит тему, а не только стиль.

Логистическая регрессия

Как работает

• Для каждого класса считает линейную оценку: сумма «признак × вес» плюс сдвиг.

• Преобразует оценки в вероятности через softmax: они неотрицательны и суммируются в 1.

• Обучение минимизирует перекрёстную энтропию — штраф за низкую вероятность истинного класса.

• L2-регуляризация удерживает веса малыми; параметр C задаёт силу регуляризации.

• class_weight='balanced' повышает вес редких классов.

Что важно

• Плюсы: устойчива при большом числе признаков, даёт вероятности, коэффициенты интерпретируемы.

• Минусы: только линейные границы; при сильной корреляции признаков веса нестабильны.

• В работе — основная линейная базовая модель.

• Сравнение C: 0,03–10; балансировка классов критична.

SVM, деревья и ансамбли

Идеи моделей

• Линейный SVM ищет границу с максимальным зазором; C штрафует ошибки.

• SVM не даёт вероятностей: только решающую функцию; тройка лучших считается по её значениям.

• Случайный лес: много деревьев, каждое на своей подвыборке; голосование.

• Сверхслучайные деревья (ExtraTrees): пороги в узлах выбираются случайно; меньше дисперсии, быстрее.

• Мягкое голосование усредняет вероятности нескольких моделей.

• Стековое объединение обучает мета-классификатор на предсказаниях базовых моделей.

Сильные и слабые стороны

• Деревья ловят нелинейные взаимодействия: «если многоточий и короткая реплика — это Jas».

• Легко переобучаются на 2005 объектах; нужны ограничения глубины и листьев.

• Дельта Барроуза: расстояние до центроида по частотным словам — классическая проверка.

• Ансамбль полезен, если ошибки моделей не полностью совпадают.

• Стековое объединение в наших данных переобучилось и оказалось хуже голосования.

Как проверяем

Схема

• Стратифицированная перекрёстная проверка: 5 блоков, доли классов сохраняются.

• Повторяем 3–5 раз с разными разбиениями — оцениваем разброс.

• Вложенная проверка: внешние блоки — оценка, внутренние — настройка.

• Отложенная выборка 80/20 — независимая проверка финальных кандидатов.

Контроль утечек

• TF-IDF, SVD, масштабирование и отбор признаков обучаются только на обучающей выборке.

• Тестовая выборка не участвует в выборе гиперпараметров или признаков.

• Проверка перестановкой меток дала 0,027–0,028 — случайный уровень.

• Почти-дубликаты внутри одного класса не пересекают обучение и тест.

• Метрики: макро-F1, сбалансированная доля, тройка лучших, MCC, каппа, матрица ошибок.

Метрики: что и зачем

Метрика	Что показывает
Макро-F1	Средняя F-мера по 34 классам; все персонажи равны
Доля правильных	Доля совпавших ответов; обманчива при дисбалансе
Сбалансированная доля	Средняя полнота по классам
Тройка лучших	Истинный класс в тройке лучших
MCC / каппа	Согласие с поправкой на случай
Матрица ошибок	Кто с кем путается

Почему не доля правильных ответов

- Случайный уровень при 34 классах: macro-F1 ≈ 0,029, top-3 ≈ 0,088.
- Модель, всегда выбирающая Abigail: доля правильных ответов ≈ 6,9%, macro-F1 ≈ 0,004.
- Поэтому основной вывод делаем по macro-F1 и доверительным интервалам.
- Дополнительно смотрим F-меру по каждому персонажу и матрицу ошибок.

Блок 3

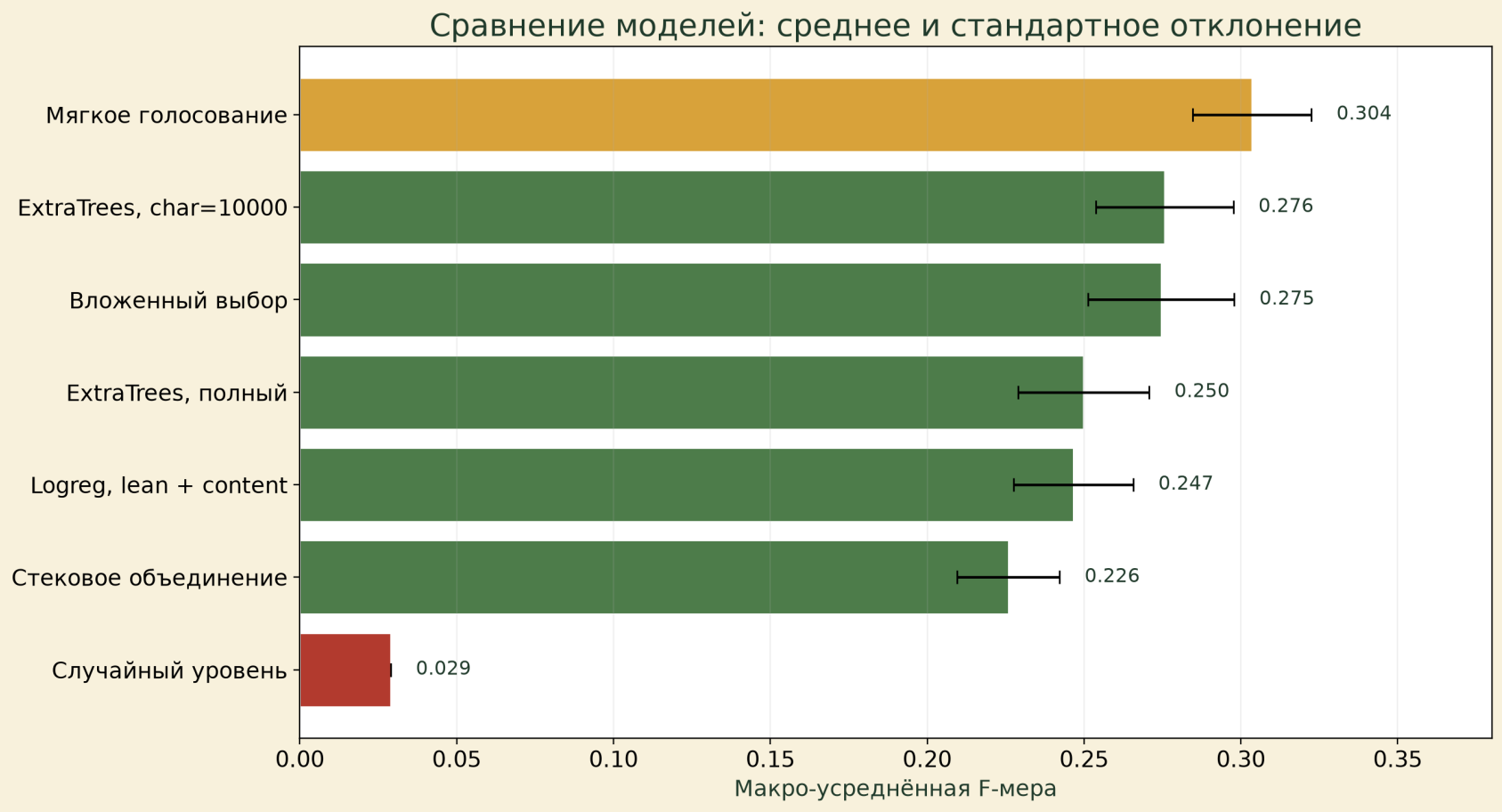
Результаты и интерпретация

Что получилось, кого модель путает и что это значит

- Лучшие модели и масштаб качества
- Стиль или содержание
- Матрицы ошибок и пары персонажей
- Ограничения и выводы

Рассказчик 3

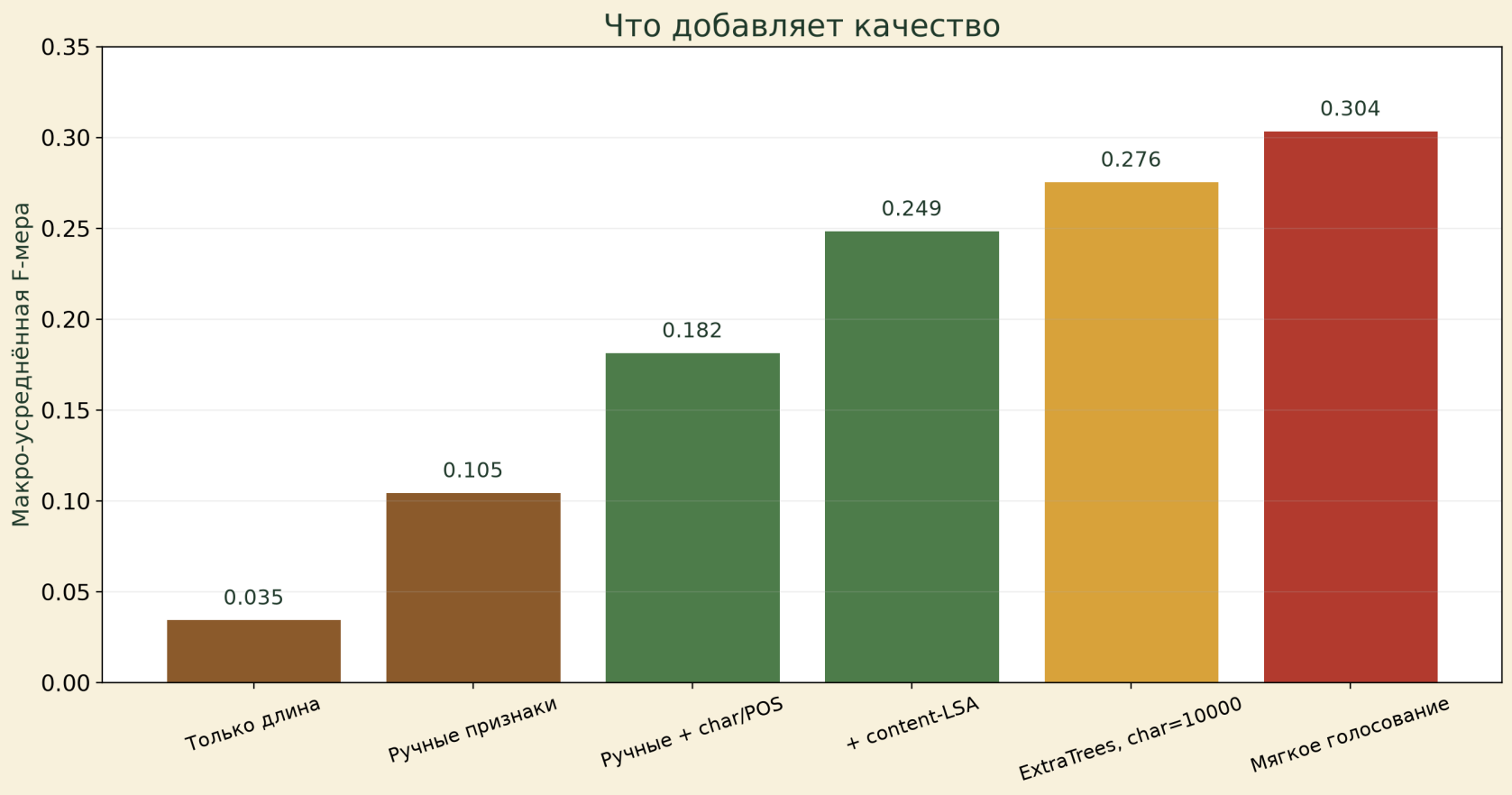
Главный результат



Макро-F1

- Мягкое голосование: 0,304.
- ExtraTrees, char=10000: 0,276.
- Вложенный выбор: 0,275.
- Случайный уровень: 0,029.
- На отложенной выборке: 0,333.

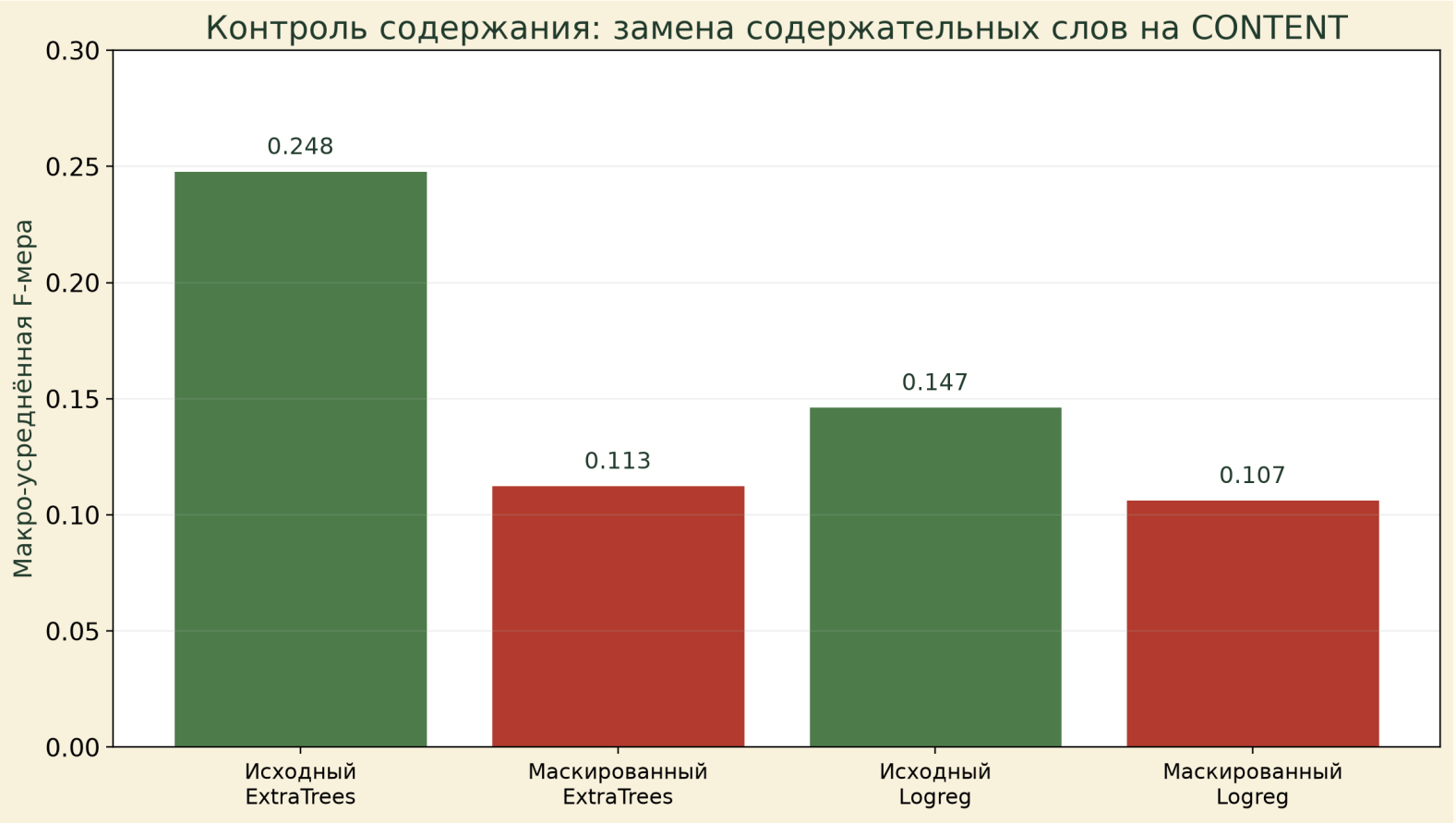
Что добавляет качество



По шагам

- Ручные признаки: 0,105.
- + char/POS n-граммы: 0,182.
- + content-LSA: 0,249.
- ExtraTrees + char=10000: 0,276.
- Мягкое голосование: 0,304.
- Основной прирост — лексика и n-граммы, не «чистый синтаксис».

Стиль или содержание?



Контроль

- Исходный текст: 0,248.
- После замены содержательных слов: 0,113.
- Падение более чем вдвое.
- Остаток 0,113 всё ещё в 3,9 раза выше случайного.
- Значит, стиль есть, но содержание доминирует.

Как читать матрицу ошибок

Маленький пример

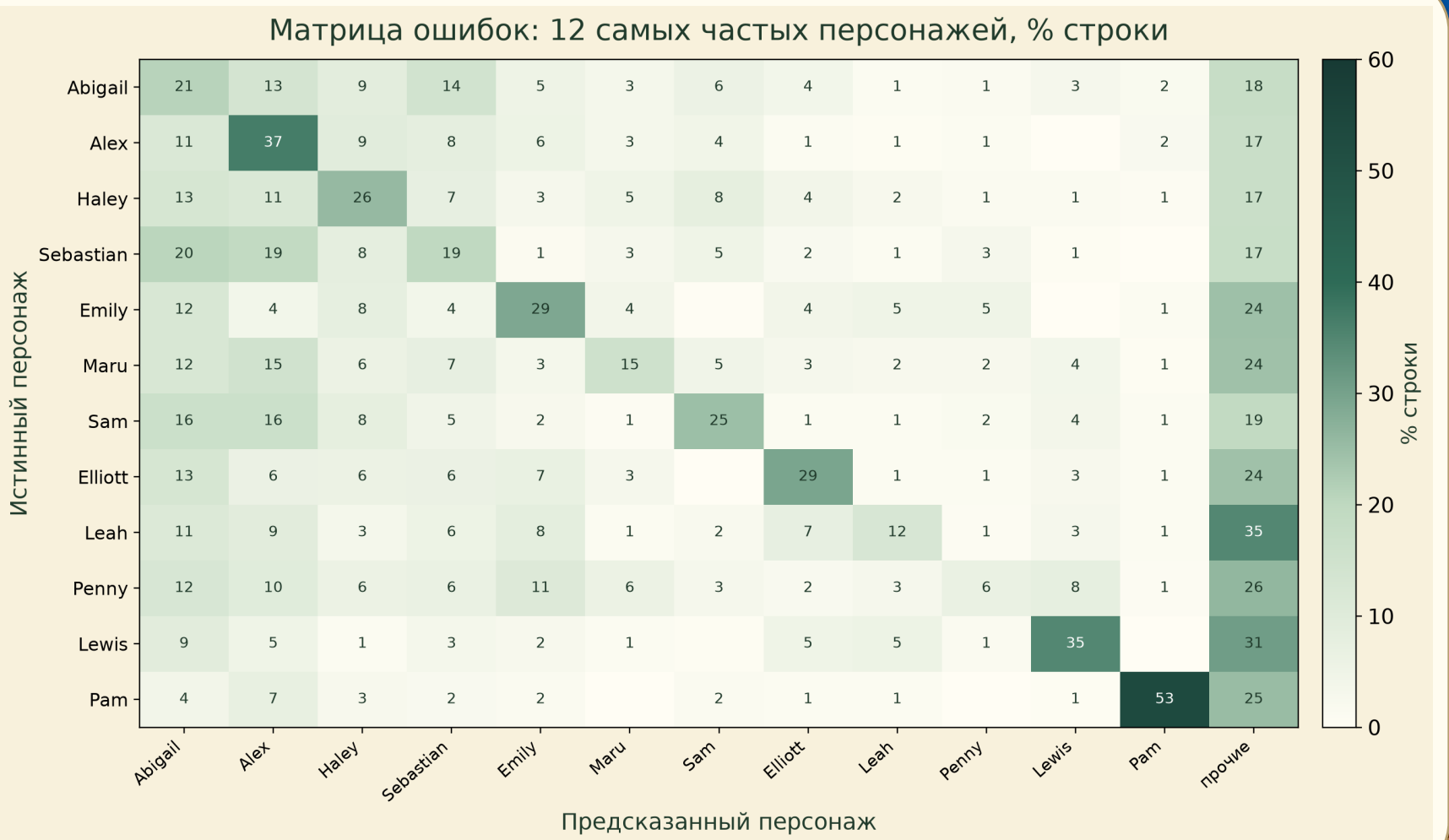
истина \ прогноз	A	B	C
A	70	20	10
B	15	75	10
C	5	25	70

Строка — истинный персонаж. Столбец — предсказание модели. Диагональ — правильные ответы. Вне диагонали — ошибки. Числа — проценты от строки.

Как показываем

- Полная матрица 34×34 на доске нечитаема.
- Поэтому на слайде — 12 самых частых персонажей и столбец «прочие».
- Отдельно показываем 18 самых частых пар ошибок.
- Полная матрица 34×34 — в раздатке/приложении.
- На доске смотрим диагональ, крупные внедиагональные клетки и «прочие».

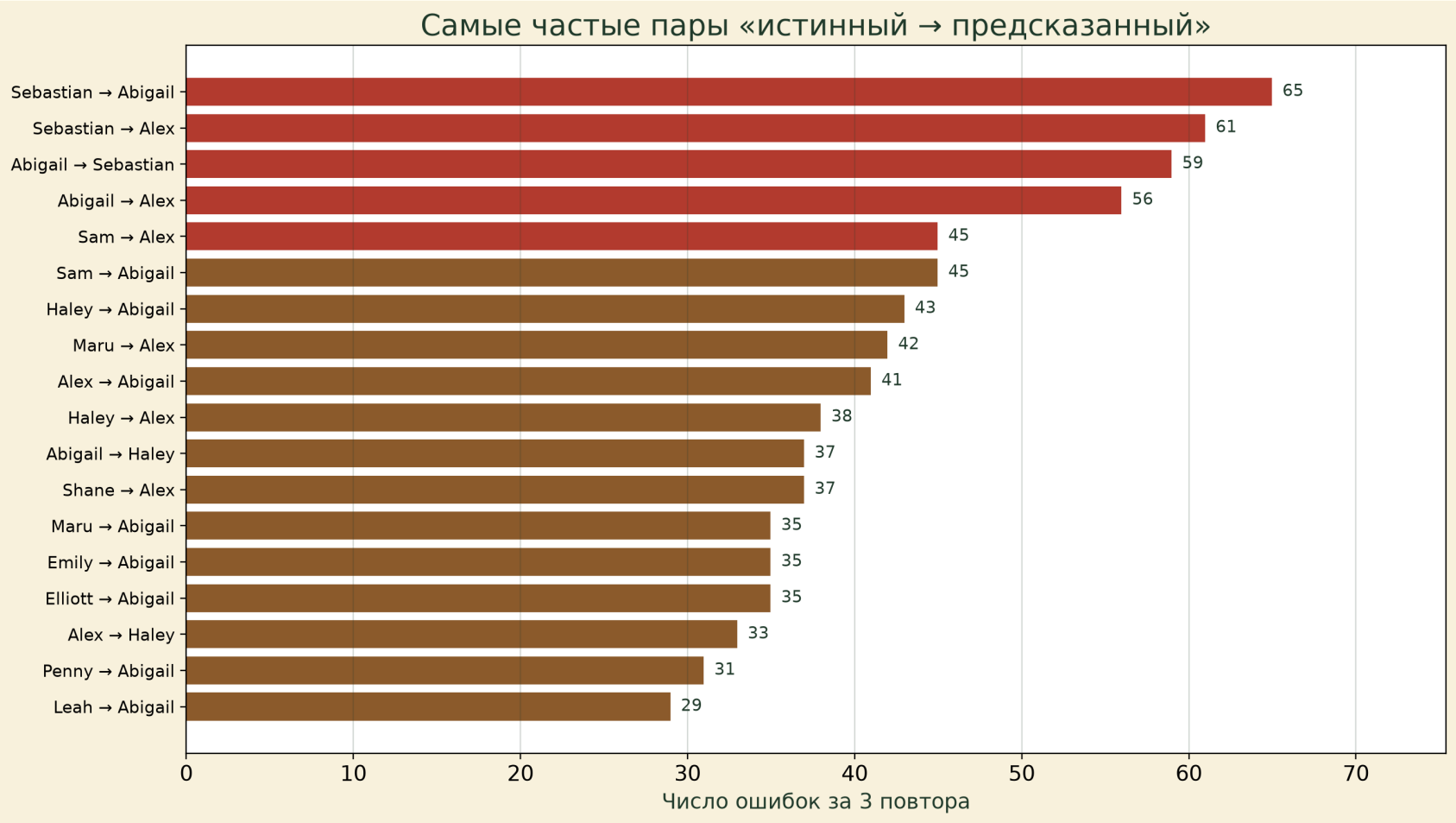
Матрица ошибок: 12 персонажей



Что видно

- Диагональ: доля правильных ответов внутри класса.
- Abigail: 21% верно, 18% уходит в «прочие».
- Alex: 37% верно, 11% → Abigail.
- Sebastian: 19% верно, 20% → Abigail.
- У многих строк крупная доля «прочих»: модель не находит редкие классы.

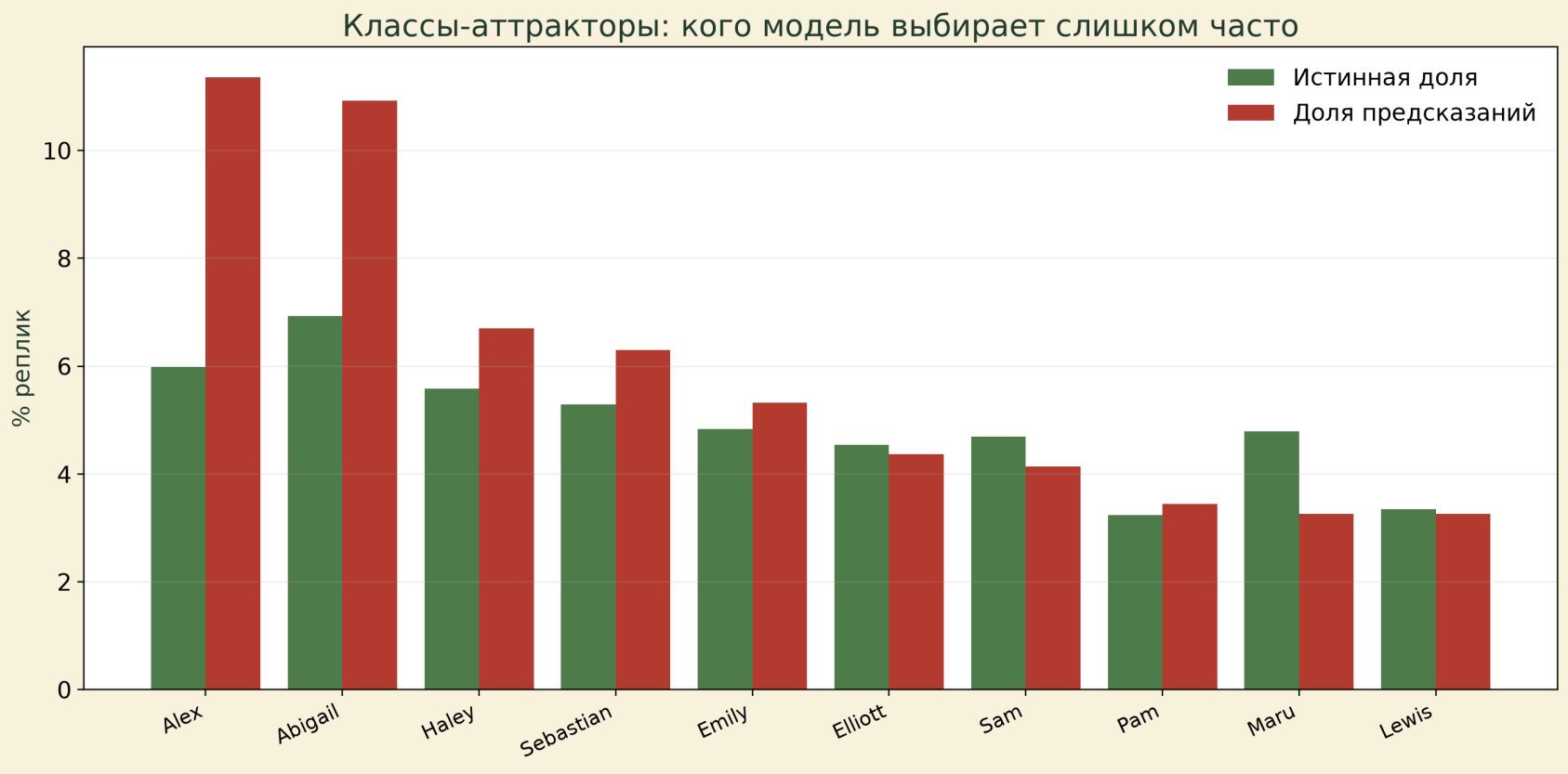
Кого модель путает



Частые пары

- Sebastian → Abigail: 65.
- Sebastian → Alex: 61.
- Abigail → Sebastian: 59.
- Abigail → Alex: 56.
- Основной клубок — подростки/молодые взрослые: Abigail, Alex, Sebastian, Haley, Sam.

Классы-аттракторы



Кто перетягивает

- Alex: 11,4% предсказаний при истинной доле 6,0%.
- Abigail: 10,9% при 6,9%.
- Редкие классы недопредставлены.
- Это не «демографическая» путаница, а стягивание к популярным голосам.
- Коррекция приоров не улучшила macro-F1.

Персонажи: кто различается



Факты

- Лучшие: Willy, Evelyn, Pierre, Leo, Pam.
- Худшие: Kent, Mister Qi, Penny, Abigail, Leah.
- Редкие классы — широкие интервалы.
- Мягкое голосование улучшает Wizard, Dwarf, Vincent.

Почему персонажи путаются

Причины

• Один автор и общие сюжетные контексты: персонажи говорят об одних и тех же событиях.

• Сходные возраст и роль: подростки/молодые взрослые используют похожие темы и интонации.

• Короткие реплики: мало слов, чтобы проявить индивидуальный синтаксис.

• Дисбаланс: редкие классы реже выбираются моделью.

• Общие формулы вежливости, шутки, междометия.

Что это значит

• Клубок: Abigail, Alex, Sebastian, Haley, Sam.

• Нелюди (Krobus, Dwarf, Wizard) различаются лучше, но Wizard и Dwarf редки.

• Дети (Jas, Vincent) имеют короткие реплики; их легче узнать по длине.

• Это не доказательство сходства «характеров», а свойство текстовых признаков.

• Нужны сцены и больше данных, чтобы отделить стиль от контекста.

Что означают оценки

Уточнение интерпретации

- 0,30 макро-F1 при 34 классах — это примерно в 10 раз выше случайного уровня 0,029.
- Абсолютное значение умеренное: по одной короткой реплике персонажа действительно трудно определить.
- Это не означает, что автор «не старался» или что стиль неразличим. Наоборот, устойчивое превышение случайного уровня говорит о намеренно различных голосах.
- Умеренность объясняется короткими текстами, дисбалансом классов, общими темами и формулами, а не отсутствием стиля.
- Мы не решаем задачу «он или не он писал»: все реплики одного автора. Мы решаем, какой персонаж говорит.

Ограничения

Данные

- Один корпус, одна игра, один автор.
- Нет внешней валидации на другом корпусе.
- Короткие реплики и дисбаланс классов.
- Ручная фильтрация корпуса не воспроизводится.
- Нет идентификаторов сцен и диалогов.

Методология

- Выбор признаков и ансамбля опирался на предыдущие результаты.
- Вложенная проверка снижает, но не устраняет оптимизм.
- Содержательные слова и тема вносят большой вклад.
- Простая калибровка приоров не помогла.
- Стековое объединение переобучается на малом числе объектов.

Выводы

Пять тезисов

• Персонажи статистически различимы: лучшая модель — 0,304 макро-F1, консервативная оценка — 0,275 при случайном уровне 0,029.

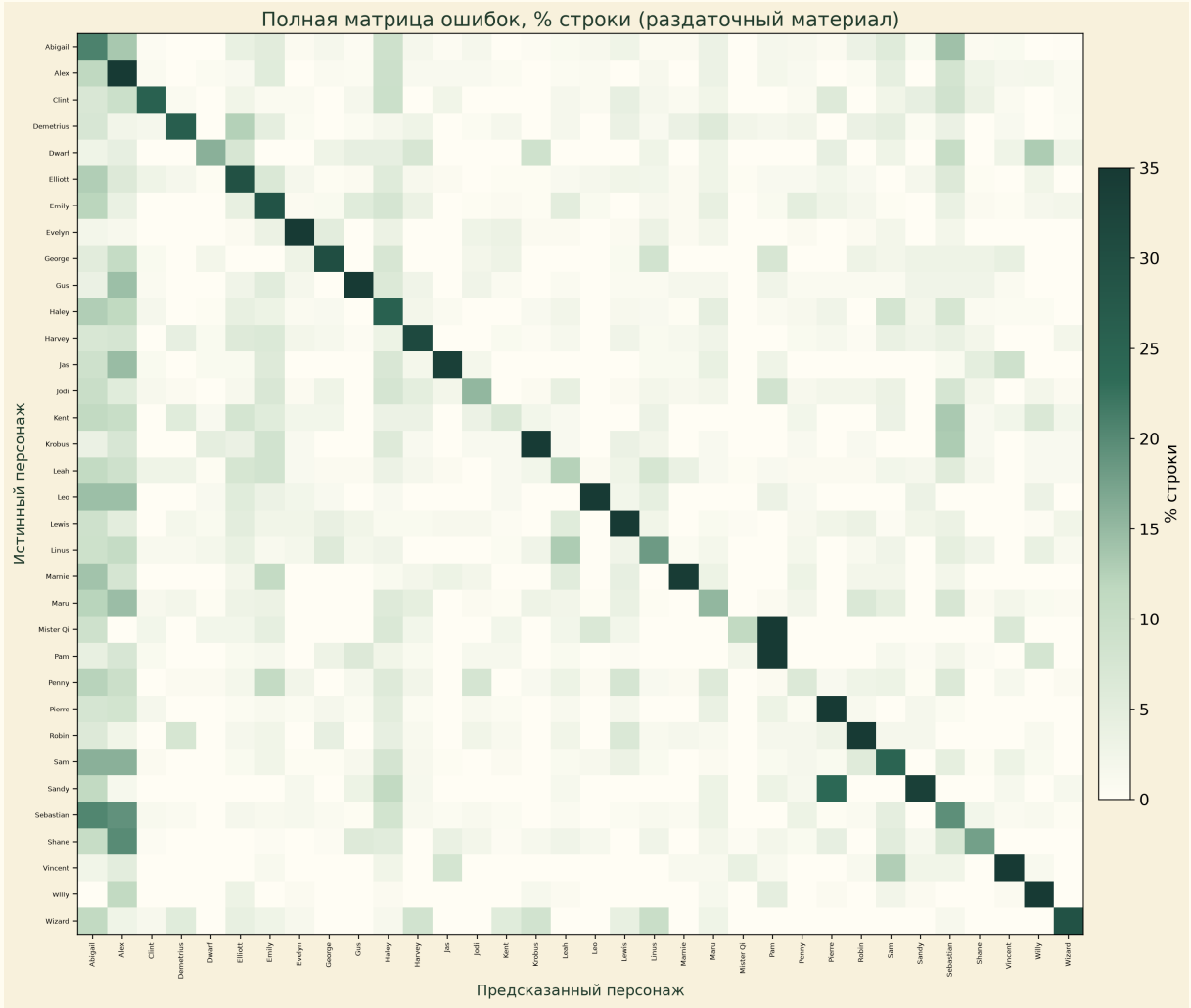
• Основной вклад дают лексика, символьные n-граммы и LSA по содержательным словам; функциональные слова и синтаксис сами по себе слабы.

• После замены содержательных слов качество падает с 0,248 до 0,113: содержание доминирует, но стилевой сигнал сохраняется.

• Главная проблема — редкие классы и классы-аттракторы; мягкое голосование помогает, но не решает её полностью.

• Корректный вывод: различимы речевые маски по лексико-тематическому профилю; для утверждения о «чистом стиле» нужны внешняя валидация и тематический контроль.

Приложение: полная матрица 34×34



Как использовать

- Полная матрица 34×34 — для раздаточного материала.
- На проекторе она слишком плотная; в докладе используем top-12 и top-18 пар.
- Строки — истинные персонажи; столбцы — предсказания; цвет — доля строки.
- Диагональ — правильные ответы; тёмные внедиагональные клетки — систематические ошибки.

Спасибо за внимание!

Вопросы, обсуждение и предложения по экспериментам

«Один автор — разные голоса»

